7 February 2008

DECLARATION OF WILLIAM A. HUBER, PH.D.

1. I apply mathematical and statistical methods to understand and solve environmental problems. During the last 20 years I have worked full time in this capacity as a consultant. My clients have included approximately 400 corporations, government agencies, nonprofit organizations, and attorneys throughout the US and worldwide.

2. In 1978, Haverford College awarded me a B.A. in philosophy and mathematics with high honors. In 1985 Columbia University conferred a Ph.D. in mathematics. Both degrees required demonstrated mastery of linear algebra, the mathematics underlying principal component and factor analysis. My Ph.D. thesis was in Lie algebras, a generalization of linear algebra.

3. My work experience includes basic research at Oak Ridge National Laboratories in chemistry and physics; teaching and research in departments of mathematics, engineering, and geology at distinguished colleges and universities; development of commercial software; environmental consulting; statistical consulting; and developing new methods of statistical and spatial analysis. The environmental consulting has focused on designing investigations of liquid and solid media (surface water, groundwater, soils, sediments, and industrial wastes) and using statistical methods to evaluate, understand, and communicate the results. My published, peer-reviewed papers address environmental site investigation, environmental statistics, image analysis, public health statistics, risk assessment, geography, and atomic physics. (For some of these papers I relied on the *Systat* package of statistical and graphical software, which I began using 19 years ago.) I have developed and taught professional and graduate-level courses in geographic information systems, statistics for groundwater monitoring, designing environmental investigation programs, and environmental statistics, as well as undergraduate courses in mathematics, statistics, computer science, and exploratory data analysis. I have served by invitation on peer review panels for the US EPA concerning probabilistic risk assessment, disinfection byproducts in water supplies, and groundwater monitoring. I have applied for US

patents for new methods of spatial and demographic analysis. Commercial software that I have designed and developed, which includes more than 50 products, provides capabilities to manage groundwater and surface water monitoring data, perform geostatistical analysis, evaluate environmental sampling data, design environmental sampling programs, and visualize geospatial data. A current curriculum vitae is attached.

4. Defendants, through their counsel, retained me in January 2008 to evaluate the production and testimony of Roger Olsen, expert for Plaintiff, specifically with regard to the details and proper interpretation of the principal components analysis (PCA) he performed and the subsequent poultry litter "signature" he derives from it. Dr. Olsen uses PCA as the basis of a "factor analysis" (a term which he eschews in his deposition, although it is appropriate in this context). He interprets PCA outputs as "factors" or "signatures" associated with alleged causes of environmental contamination within the Illinois River Watershed.

5. PCA is a well-known method of analyzing multivariate data, such as when numerical measurements of multiple analytical parameters are performed on samples of environmental media. PCA, like any other multivariate statistical method, is complicated. Except in the simplest cases it is impossible to perform without software. Its use is beset with traps and pitfalls for the unwary or inexperienced. Many choices and much judgment go into it, resulting in a wide variety of possible output. The results can be difficult to interpret. In the hands of a knowledgeable practitioner, PCA can provide insight into correlations among the variables. However, PCA when applied to water, sediment, or soil data does not by itself pinpoint "sources:" it requires subjective interpretation, known as "reification." (John Davis, author of a seminal and popular text on PCA, writes, "Possibly some analysts feel the use of this term [reification] makes this subjective process more respectable." [Davis, JC, *Statistics and Data Analysis in Geology*, 2nd Ed., 1986. At page 536.]) To drive this point home, Davis characterizes factor analysis (including PCA) as "a controversial and poorly understood methodology that extends the beguiling promise

of instant insight to the researcher faced with more data than comprehension." [*Ibid.*, p. 516.]

6. It is useful to distinguish mathematical techniques from statistics from science. As a mathematical technique, PCA "is a method of decomposing a correlation or covariance matrix[1]." [Stenson H. and L. Wilkinson, *Factor Analysis*, in Chapter 12 of the Systat 12 manual "Statistics I II III IV," 2007.] As such, when presented with valid input, PCA will *always* produce results in the form of the "signature" Dr. Olsen identifies and relies on. As a statistical technique, PCA is *exploratory*. This means it is not intended (and rarely used) for testing hypotheses or conferring statistical significance on conclusions. The credibility of PCA results depends on conducting preliminary data evaluations. Normally these would accompany any well-conducted statistical study[2]. These evaluations would normally and routinely include graphical and numerical assessments. A "graphical" assessment is a meaningful picture of the numbers. It can be as simple as plotting one value against another on a sheet of paper. A "numerical" assessment, or "summary statistic," is what we usually associate with statistics: an average, a range, a correlation coefficient, or some other calculation intended to describe the data. Graphics are essential: it is well-known they can reveal more than a bunch of summary statistics, because the graphics *show* the data, often in great detail. The assessments one would expect to precede a PCA include (1) looking at the individual measurements (analytical parameters) in the dataset, (2) looking at every possible pair of measurements together, (3) breaking the data into groups to assess the influence of other variables (such as the medium sampled, the time of sampling, sampling conditions, and so on), and (4) consideration

[1] "Correlation coefficients" and "covariances" are numbers that reflect the associations between two variables. When many variables are present—Dr. Olsen testifies to using 25 at once—one can compute such a number for each pair of variables. A "matrix" is simply an orderly tabulation of all these numbers.

[2] This might not be apparent to some scientists, because the data evaluations described here, known as "exploratory data analysis" (EDA) often are not published. But nevertheless textbooks and abundant statistical literature, dating back over 30 years, testify to the importance of doing EDA. Good scientific papers will at least briefly summarize EDA results. Modern statistical software makes it easy to produce the exploratory graphics and statistical summaries, but one still has to produce them, look at them, and take the actions they indicate.

of appropriate data weighting. (This last point deserves further explanation. Data are uncertain: two samples of the same thing, taken at the same time, will yield measurements that differ at least slightly. Large differences produce large uncertainty. Investigators manage this measurement uncertainty in many ways, such as by repeating the measurements, taking multiple samples per location, changing the intensity of sampling over space and time, using samples of larger volume, compositing samples, and so on. The data analyst must therefore consider the different amounts of uncertainty present and, if necessary, weight the data accordingly.) Finally, interpreting the results of a PCA is not purely a matter of statistics: it requires a framework of scientific theory and hypotheses.

7. Dr. Olsen's affidavit of 26 October 2007 provides no specific information about his PCA, which is the basis for opinions 6, 7, 8, 9, and 10 therein. With the aim of understanding and reproducing his analyses, I have reviewed his affidavit, preliminary transcripts of his testimony of 2 February 2008, portions of exhibits 00000011 through 00028613 (3413 computer files, many of them compressed archives of multiple files, comprising 1.42 gigabytes), and additional computer files in folders named "Data Base 7" and "Data Base 8." I expected to find, at a minimum, documentation that is standard and necessary for any statistical study: namely, exactly how the data were processed, what calculations were performed, what the results of those calculations were, and how those results were interpreted. This material includes many undated versions of Dr. Olsen's database of investigation results and various spreadsheets that appear to be the output of PCA calculations. As far as I can ascertain, there exists little or no documentation concerning the uses of these files, their interrelationships, or the sequences of commands needed to perform PCA calculations with the statistical software Dr. Olsen used, *Systat*. When Dr. Olsen was asked to provide such information during his testimony, he often failed to respond with meaningful answers. When he did answer, many times his responses were confused, reflected poor statistical practices, or contradicted the evidence in the aforementioned files.

8. Using these materials, and by means of extensive experimentation, I have been able to closely approximate some of Dr. Olsen's recent results contained in spreadsheets from the "Data Base 8" folder dated January 2008. This appears to be his most recent work. I believe that the small differences between my results and Dr. Olsen's are due to slightly different selections of the data and not due to different statistical procedures. It is clear that he selects groups of samples, selects samples from those groups, and also selects a set of analytical parameters. Numbers in the "EDA_Value" field of the database are averaged, most likely by location (so that he can assign "factor scores[3]" to locations in order to map them). Logarithms[4] of the results are computed. (It is possible the logarithms are computed first and then averaged[5].) Pairwise[6] correlation coefficients in the resulting "cross table[7]" are computed, the PCA mathematical procedures are performed, additional factor analysis calculations called "rotations[8]" are computed, and the final output is collected within a single spreadsheet. Excel (spreadsheet software) and Systat (the statistical software) appear to be used for all the computations and data manipulation. The intervening calculations, graphics, and output appear to be deleted or lost. There is no record of the commands sent to Systat to control its calculations and specify its output. Only

[3] The principal output of PCA is a set of "factors" or "components." Each one looks like a set of measurements, one number for each analyte. The factors have no intrinsic meaning, but they can be used mathematically to re-express *any* set of measurements as a weighted sum of factors, at least approximately. The "score" for a factor is the weight that appears in this re-expression.

[4] Logarithms were invented to simplify the arithmetic of multiplication. They are not used for that purpose here. Environmental data tend to extremes: a few very large values often occur among many smaller ones. To prevent the largest values from unduly influencing the results, it is standard practice to re-express the numbers. The logarithm is one possible means of re-expression that reduces the spread among the largest values and increases the spread among the smallest, thereby "balancing" the distribution in a beneficial way.

[5] The sequence of these operations matters. However, the effects on the PCA in the examples I examined were so small that I could not determine whether Dr. Olsen averaged first or took logarithms first.

[6] There is a technical distinction between "pairwise" and "listwise" computation of correlation coefficients. The difference is important when data are missing, which is the case in Dr. Olsen's data. The two methods do give different results.

[7] This non-standard term appears in some of Dr. Olsen's spreadsheets. It refers to the database of averages (or their logarithms).

[8] Fortunately, it is not necessary to explain what "rotations" are, because Dr. Olsen has testified that he does not use these results.

three spreadsheets remain: one contains data, another contains obscure lists intended to document which data were selected, and the third contains a table of "loadings" for five "factors." There does not seem to be any record of the procedure used initially to extract some data from the investigation database (or whether a single, consistent investigation database exists at all).

9. Dr. Olsen's procedure incorporates a fundamental, widely-recognized error known as the "ecological fallacy." The ecological fallacy consists in making inferences about individuals—in this case, the chemical and biological constituents in samples of surface water, groundwater, sediments, and soils—based on their aggregate behavior. Dr. Olsen's aggregates include averages of measurements obtained at different times and under varying conditions, such as high flow, normal flow, and base flow at the same location. It is a mathematical fact that correlations of averages will be stronger than correlations among the individuals. This causes the PCA to produce incorrect "signatures" (in Dr. Olsen's terminology) that falsely associate constituents with each other and create incorrect scores at the locations. One reason I am convinced this mistake occurred in Olsen's work is that repeating the procedure that emulated his results, but applying it to *individual* measurements, produces distinctly different results. Moreover, the effect of the aggregation is not even uniform. The number of samples at each location varies. There are often repeated measurements of some analytes. Thus, for a given constituent, there may be just one value or many dozens of values contributing to its average. This should cause the analyst to assign varying weights to the averages for the PCA computations, but there is no sign that any kind of weighting was calculated or used.

10. A second fundamental error is failure to divide the data into appropriate groups for analysis. Throwing all types of data together into a single PCA—edge-of-field samples, high flow samples, base flow samples, and so on—results in correlations that are likely incorrect for *any* meaningful subset of the data. (On a farm with equal numbers of chickens and cows, merging data in this manner would lead to the conclusion that the animals are one-winged tripeds. Mathematically correct, yes; but

meaningful, no.) For example, when one performs a PCA on appropriate subgroups of the surface water data, as defined in the database by Dr. Olsen himself, dramatically different "signatures" result[9]. They do not indicate any pervasive, watershed-wide signature; in particular, there is no evidence from this exercise that just one or two single phenomena, such as poultry litter spreading, can adequately characterize the correlations in these data. This 80 mile long, eight-county watershed containing over 2500 miles of streams and rivers interacting within "multiple compartments" is likely too rich and complex for such a simplistic characterization.

11. A third fundamental error is to interpret correlation as causality, as Dr. Olsen has testified repeatedly. PCA relies solely on information about correlation. No matter how many variables it uses, it is still just correlation. Making additional measurements of one's samples cannot magically turn these correlation coefficients into an underlying cause. In some cases, as in a randomized controlled study, a weak correlation among two variables would be evidence of a common underlying cause, and in other instances (notably economic analyses) a strong correlation among thousands of variables may reflect nothing more significant than the passage of time or some other phenomenon with which all the variables are naturally (and trivially) associated[10]. In particular, water quality concentrations tend to increase and decrease in lockstep with environmental factors that influence groups of chemically or biologically similar constituents in common: amount of flow (or dilution),

---

[9] These signatures are not likely to be valid, due to the other problems inherent in Dr. Olsen's PCA procedure. The point is that they differ from each other and from the signature he derived.

[10] An analogy with a more familiar situation may clarify this point. Imagine a hypothetical study of blood lead in Oklahoma City children. Investigators evaluate a set of randomly chosen children. They give them a battery of tests to measure knowledge and aptitude. They also measure bodies: head circumference, shoe size, height, weight, and so on. And of course they measure lead in the blood. *All* of these variables will be correlated, because all of them tend to increase with age. The first component in a PCA of these data will have "high loadings" on blood lead and on most of the other variables as well. It would be ludicrous, though, to conclude knowledge, aptitude, and body size are *caused* by the lead a child has absorbed!

In this analogy, the children of Oklahoma City correspond to the Illinois River Watershed, the body measurements correspond to biological measurements of water samples, the test results correspond to chemical measurements of water samples, and the blood lead measurements correspond to the analytes believed to be associated with chicken litter.

precipitation, sediment loading, time of day, season, *etc*. Therefore, before one ever collects a single sample in a study like this, it is to be expected that the first (highest) component of any PCA analysis will reflect the combined influences of such factors and therefore have "high loadings" on (strong correlations with) most of the variables in the study. Indeed, this is exactly what the output in Dr. Olsen's most recent PCA spreadsheets contains, and it is primarily on the first component that Dr. Olsen relies for his opinions. As the preceding footnote points out, having lots of variables does not necessarily strengthen the results.

12. Dr. Olsen's testimony indicates he is relying solely on "factor loadings" in a single component to identify his poultry litter "signature." This is a limited and incorrect means of interpreting the output. The Systat manual [*op. cit.*] warns about this. "Usually these loadings are not useful for interpreting the factors." [At page I-472.] It is important to understand that PCA, as Dr. Olsen performed it, works with correlations only, and therefore does not directly reflect the actual concentrations and bacterial counts in the data. Thus, the "loading" of a "factor" on a single variable says nothing about the importance of that variable or its potential effects on water quality. Davis also warns, "In circumstances [where correlations are used in PCA], we must remember that a property that seems relatively insignificant may exert a strong influence on the analysis." [*Op*. *cit*., p. 536.]

13. There are signs that Dr. Olsen's procedures for handling data introduced errors before any analyses were performed. Evidence for this includes the fact that multiple, conflicting locational coordinates (latitude and longitude) appear within individual spreadsheets he used for the PCA. Some of the various coordinates assigned to a given location differ by miles, others by hundreds of feet. This would result in plotting the results at the wrong locations, potentially changing their interpretation. If such important inconsistencies occur for coordinates, one must suspect that substantial errors occur in the other fields, too, including sample identifiers, group identifiers, and the values themselves. The frequency of locational errors, which affect around ten percent of the locations in one recent spreadsheet, is alarming. If a

similar error frequency exists for the rest of the data, then the PCA results could be entirely incorrect.

14. Dr. Olsen's testimony about handling nondetects is consistent with the evidence in his computer files: he replaces such values by their detection limits. Many rules of thumb exist in the environmental literature for treating nondetects, but virtually all of them are intended for other purposes: namely, estimating average concentrations. For computing correlations, Dr. Olsen's method is inferior, because it confounds variation in observed quantities (concentrations or bacterial counts) with variation in laboratory reporting limits (which reflect many things that are not associated with concentrations at all). This can create apparent correlation where none exists or underestimate correlation that does exist. Taking logarithms of the data, as it seems Dr. Olsen eventually did, exaggerates these effects. There are better methods to estimate correlations when data are missing or "censored[11]" by nondetects.

15. Dr. Olsen appears to ignore the fact that some bacterial counts are censored on the right ("greater-than values"). This can have consequences similar to treating nondetects as values equal to the detection limits: that is, the correlation coefficients involving bacterial counts may be biased, creating incorrect PCA "signatures."

16. The investigation database suffers from incompleteness: many of the samples Dr. Olsen relies on do not have measurements of all the parameters he uses. (Completeness—obtaining a sufficient proportion of the measurements needed to support intended data analyses—is one of the US EPA's data quality objectives.) All the evidence in Olsen's computer files indicates he computed correlations in a "pairwise" manner, both in the spreadsheets and in the statistical software. When the pattern of missing data happens to be correlated with the results themselves—which

[11] "Censoring" is a statistical term describing data whose values have been cut off by a threshold like a detection limit. Censoring is a particular problem for environmental data analysis, where nondetects are the norm. About 25 years ago researchers began developing effective methods to incorporate censored data in their analyses without biasing the results. Ten years ago many of these methods started to become more accessible by appearing in textbooks and regulatory guidance. Unfortunately, the older guidance is still widely used and many older environmental scientists are probably unaware of the newer methods.

is likely—pairwise computation will bias the correlation coefficients and thus invalidate the PCA results. There is no evidence that Dr. Olsen performed the necessary tests to evaluate and correct this problem.

17. There is little evidence that Dr. Olsen conducted most of the routine evaluations, tests, sensitivity analyses, and other data procedures needed to assure a reliable PCA. (In fact, by automating the interaction with the statistical package, the spreadsheet software he created to conduct PCA makes these procedures inaccessible to the user.) Many of these auxiliary procedures are graphical. It is well established that responsible, accurate evaluation of data must include graphical displays: "above all, draw a picture!" has been the watchword of the foremost statisticians and data analysts. The graphics one would expect to see, either as computer files or as files of commands to produce the graphics, include probability plots, scatterplot matrices, scree plots, and factor loading plots. Dr. Olsen's testimony suggests he may have looked at probability plots and I have found one scatterplot matrix among the computer files he produced. However, he has testified that the work on which he is relying includes *no* graphical evaluations at all.

18. Some aspects of Dr. Olsen's PCA methodology are nonstandard and likely to produce misleading results. In particular, selecting five factors at the outset (rather than letting the PCA results establish the proper number of factors to examine) is erroneous. In some cases five factors is too many (so that anything Dr. Olsen sees in the last few factors is likely just random "noise") and in other cases it is too few (so that Dr. Olsen will fail to identify "signatures" that can distinguish various chemical behaviors).

19. In summary, Dr. Olsen has not adequately described or documented his principal component analyses [paragraph 7]. Despite this, I have been able to reconstruct the procedures he most likely used [¶ 8]. They incorporate three fundamental errors, any one of which renders the results invalid: the ecological fallacy [¶ 9], grouping unlike data [¶ 10], and confusing correlation with causation [¶ 11]. He does not interpret the results properly [¶ 12]. Various additional technical deficiencies are apparent [¶¶ 13

– 18]. Assuming, hypothetically, that Dr. Olsen's PCA results were correct, they do not support his conclusion that there is a pervasive, watershed-wide "signature" revealing the presence or effects of poultry litter.

20. To the extent Dr. Olsen's work is based only on the information provided to me, he does not have an adequate foundation for his principal component analysis or the conclusions he draws from that. To the extent there is additional documentation that emerges, I reserve the right to amend my opinions.

Signature: William A. Huber

William A. Huber, Ph.D.

Date: 7 February 2008 ____________

My fee for this work conforms to the QUANTITATIVE DECISIONS Standard Schedule of Charges for 2008, attached.

During the last four years (calendar years 2004 through 2007) I have been deposed as an expert in one case, Hammerwood Avenue, L.P. (Abrams Investment Co.) *vs*. Thermo Electron Corporation, *et al.*, United States District Court, Northern District of California, San Jose Division, Case No. CV-04-01081-JW (2006).

Quantitative Decisions

1235 Wendover Road
Rosemont, Pennsylvania 19010
(610) 527-3599
whuber@QuantDec.com

Standard Schedule of Charges
2008

Type	Price	Units	Increment
Consulting, principal	$170	per hour	¼ hour
GIS design and programming	$120	per hour	¼ hour
GIS and data management	$90	per hour	¼ hour
Travel by car	$0.43	per mile	1 mile
Direct expenses	At cost		
Other job-related materials	At cost		
Subcontractors	At cost		

Explanations

(1) Labor and expenses are itemized on invoices.

(2) Invoices are sent monthly and are payable within 30 days.

(3) “Other materials” includes equipment, software, etc. procured on our clients’ behalf.

(4) Subcontracting costs are passed on directly with no markup.

(5) There are no markups or extra charges for litigation support.

(6) Discounts up to 50% are available for not-for-profit and government organizations.

(7) Rates shown in this table are subject to change after the first of each year. Current clients will be notified in advance of any changes.

Quantitative Decisions

1235 Wendover Road, Suite 100
Rosemont, Pennsylvania 19010
(610) 527-3599
whuber@QuantDec.com
http://www.quantdec.com

WILLIAM A. HUBER, PH.D.
PRINCIPAL

EXPERTISE

Statistical analysis
Mathematical modeling
Geographic modeling and analysis
Geographic information systems (GIS)
Environmental statistics and geostatistics
Decision analysis

EXPERIENCE

Environmental: Assessment and interpretation of environmental data. Development of strategies to evaluate and improve the value of environmentally impaired sites. Development, application, and dissemination of improved methods to sample environmental media, assess the quality of data, interpret data, make optimal data-based decisions, present conclusions, and evaluate other interpretations. Negotiation and presentation. Peer review and strategy development for environmental investigations, remediation, and closure. Litigation support. RCRA groundwater monitoring statistics and water quality monitoring design.

General: Development and application of mathematical and statistical models to analyze and process spatial data, including transportation (network) analysis and real estate market analysis. Geographic Information Systems development and analysis. Statistical consulting, computer programming, and database management. Teaching. Theoretical and applied research in mathematics, statistics, and physics.

Dr. Huber has completed over 150 projects for Quantitative Decisions since 1997, including

Environmental liability assessment

Evaluation of offsite liabilities at a former pigments manufacturing plant. Developed an offsite investigation work plan, evaluated all data, supported the defense of civil and criminal claims, and provided improved methodology to the health authorities for conducting a community survey and blood sampling program (Mexico, 2002).

Assessment of potential environmental costs for brownfields redevelopment at a former refinery. Identified regions most suitable for initial development and evaluated the extent of potential soils contamination (East Coast US, 2001).

Estimation of financial liabilities at chemical manufacturing facilities, for computing insurance cost recovery and supporting the development of investigation and remedial strategies (NJ, KY, MI).

Use of decision analysis to formulate a strategic plan to address Superfund liabilities (Puerto Rico, 1997).

Environmental Statistics

Develop alternate groundwater monitoring compliance limits (ACLs) for arsenic in groundwater at a petroleum refinery and pipeline facility, 2008.

Peer review of the US EPA Guidance on Statistical Methods for Groundwater Monitoring, 2005.

RCRA groundwater monitoring permit development for a large Midwest oil refinery, 2004: comprehensive data review, selection of monitoring wells, monitoring parameters, and statistical tests; negotiation with state and Federal regulatory agencies; creation of the written permit; and development of software to streamline permit compliance.

Peer review of the Hendry County Groundwater Flow Model for the South Florida Water Management District, 2001.

Developing and defending remedial goals and optimizing remedial designs for metals in soils at Superfund sites (NY, PA).

Evaluation of groundwater monitoring data and development of an ongoing monitoring plan at two landfills at a U.S. Army Ammunition Plant (MO, 1998).

Evaluation of soils data for investigations, waste characterization, management, and disposal (CA, CT, FL, GA, IL, IN, KY, MI, MO, NJ, NY, OH, PA, VA, WI, Canada, and The Netherlands)

Statistical plans to reduce soil and groundwater sampling costs for environmental investigations (CA, CT, DE, IL, FL, NJ, NY, PA).

Developing alternate groundwater monitoring compliance limits (ACLs) for a uranium mine regulated by the NRC (WY, 1995-7).

Consulting on RCRA groundwater monitoring issues. (AR, AZ, KS, KY, MO, NJ, OH, OK, PA, TX, VA).

Risk Assessment

Expert reviewer for the USEPA of procedures developed to establish decision-making guidelines for residual disinfectant levels in drinking water and to set maximum contaminant levels (MCLs) for disinfectant byproducts (1999).

Invited presenter at the Second Workshop on the Practical Issues in the Use of Probabilistic Risk Assessment sponsored by the USEPA and University of Florida (1999).

Participation as an invited expert in the EPA's Workshop on Selecting Input Distributions for Probabilistic Risk Assessments (NY, 1998).

Litigation Support and Negotiation

Statistical support to evaluate MTBE in public water supply wells (NY, 2006).

Evaluation of lead in an industrial building and expert review of mathematical models concerning the origin and dispersion of environmental contaminants (CA, 2005-2006).

Expert testimony, geostatistics. US Department of Justice (defendant). Evaluated a complex hydrological model formulated by plaintiffs to support a $4 billion claim for natural resources damages. Discovered and testified to fundamental flaws in the estimates of a chlorinated groundwater plume extent. The client was subsequently dropped from the case (NM, 2002).

Expert review and criticism of a complex probabilistic dose reconstruction model. Provided advice to defense counsel and helped prepare for deposing expert witnesses in hydrogeology, statistics, and risk assessment (CA, 2001-2).

Independent review of local and regional groundwater data at an MTBE contaminated wellfield on behalf of a former gas station owner. Addressed regulatory concerns about data quality (high detection limits) and geological conditions (CA, 2000-2001).

Development of a waste sampling and analysis program to help a landfill demonstrate attainment of Land Disposal Restriction (LDR) standards (OH, 1999-2001).

Second opinion, peer review, and support in deposing expert witnesses for an insurance claim litigation concerning soils contamination by heavy metals at a former rail maintenance yard (PA, 2000).

Investigation strategy development, data visualization, and geostatistical analysis to help a chemical manufacturer limit liability for extensive groundwater contamination by chlorinated solvents (CA, 1997-2000).

Statistical support to defend a client against a claim of using an incorrect statistical test for RCRA groundwater monitoring at a large hazardous waste facility (OH, 1998).

Successful criticism of a probabilistic ground water model purporting to demonstrate historical landfill releases of chromium (PA, 1997).

Decision Analysis

Development of a multiattribute valuation function to prioritize 8,000 sites according to suitability for cellular towers (NJ, 1999). http://www.quantdec.com/projects/wireless.htm

Decision support for development of an open space preservation plan, Franklin Township, NJ, 1999. http://www.quantdec.com/open.htm

Modeling

Development of new techniques for interpolating and predicting demographic data. Patent applied for, 2006.

Development of new techniques to find optimal travel costs in spatially diffuse networks. Patent applied for, 2006.

Development of new methods and software to simulate, evaluate, and predict supply and demand within regional markets. In collaboration with Fiscal Associates, Newark, DE, 2003-present. Patent applied for, 2006.

Development of new methods and implementation of software to optimize reallocation of agricultural lands. Alterra, Wageningen, The Netherlands, 2002-3.

Development of improved techniques and software for the computation and visualization of contaminant plumes from regional air sources (TNO-MEP, The Netherlands, 1999). http://www.quantdec.com/projects/ammonia.htm

The types of projects and activities previously completed include

Development of strategic management plans and financial and economic evaluations using decision theory. Applied successfully to state and federal Superfund sites, utilities management, site investigations, and remedy selection.

Review and strategic development of sampling, remediation, and closure plans for many sites across the United States in EPA Regions II, III, IV, V, VII, and IX.

Expert testimony on the interpretation of surface water and ground water data at hearings with Pennsylvania and New Jersey regulators; for manufacturing facilities.

Statistical and geostatistical ("kriging") evaluation of contaminant patterns and quantities to support human health and ecological risk assessments; for Superfund sites, mines, manufacturing facilities, chemical treatment facilities, refineries, and landfills.

Critical evaluation and analysis of draft EPA guidance, for regulated facilities; for example, see "PCBs in Pipes" at http://www.quantdec.com/Articles/pcbpipe/pcbpipe.pdf.

PREVIOUS EXPERIENCE

Senior Associate, Dames & Moore, Inc., Willow Grove, PA, 1992-1997: Project management, marketing, and firm-wide technical support for issues related to environmental statistics and information management. Provided written evaluations for approximately 200 projects world-wide and participated in about 200 proposal efforts. Served private sector clients and state government agencies.

Partner, Integrated Data Technologies, Inc. (IDT), Philadelphia, PA, 1986-1992: Developed and managed an environmental software, database, and statistical consulting business.

ACADEMIC BACKGROUND

Bryn Mawr College, Bryn Mawr, Pennsylvania: Lecturer in Geographic Information Systems, 2007.

Haverford College, Haverford, Pennsylvania: Visiting Associate Professor in the Department of Mathematics, 2005-2007. Courses include *Exploratory Data Analysis* and *Statistical Methods and Their Applications*.

Penn State University–Great Valley: Adjunct professor in the Engineering Department (1997-2004). Courses include Special Topics in Environmental Statistics; Geographic Information Systems. Supervised three Masters' theses in environmental science and engineering.

St. Joseph's University, Philadelphia, Pennsylvania: Assistant Professor of Mathematics and Computer Science (1984-86).

Ph.D., 1985; M. Phil., 1980; M.A., 1979. Mathematics, Columbia University in the City of New York.

B.A., 1978. Philosophy and Mathematics double major, Haverford College, Pennsylvania, with high honors.

- College mathematics prizes 1975, 76, 77.
- Phi Beta Kappa 1977.
- Finalist, Danforth (teaching) and NSF (research) fellowships, 1978.

CITIZENSHIP

United States

PROFESSIONAL AFFILIATIONS

Associate Editor, *Environmental and Ecological Statistics*
American Statistical Association
Mathematical Association of America

SELECTED PROFESSIONAL ACTIVITIES

Peer reviewer, *Human & Ecological Risk Assessment* (1998); *Environmental Science & Technology* (1997-2002), *Risk Assessment* (1996-99), *Risk Analysis* (2003-2008), Journal *of Hydraulic Engineering* (1994-6; 2005); *Environmental and Ecological Statistics* (1994-2000); *Geotechnical Testing Journal* (1995).

Author of over 40 publicly available software programs to perform statistical and geometric analysis and visualization of data.

ESRI (GIS) Support Center User Forums annual "MVP" Award, 2003, and numerous semi-annual awards, 2002-2007. http://support.esri.com .

Editor, ***Directions Magazine*** (http://www.directionsmag.com/), 2001. *Directions* is a Web magazine, focusing on geographic information systems, with about 75,000 viewers monthly.

Contributing Editor, ***Directions Magazine***, 1999-2000 and 2002-present.

Founder and owner of a 1000-member Internet discussion group focusing on technical issues in Geographic Information Systems (GIS), 1999.

Co-instructor, *Geographic Information Analysis: Spatial Statistics Workshop*. Wheaton College, Norton, MA, June 4 – 8, 2007. http://www.nitle.org/index.php/nitle/content/view/full/1149 .

Invited speaker on *Designing Environmental Investigations with GIS* at the *2nd Annual GIS and Public Health Day: Methods and Strategies for Enhancing Environmental Health Surveillance*. Center for Public Health Preparedeness, School of Public Health, SUNY Albany, NY, May 9 – 10, 2006. http://www.ualbanycphp.org/Events/GISDay_05_09_06/default.cfm .

Invited speaker on statistics at the National Groundwater Association's Second Theis Conference, Amelia Island, Florida, November 1999.

Invited panel member, *Workshop on Selecting Input Distributions for Probabilistic Risk Assessment*, U.S. EPA, New York City, April 21-22, 1998.

Keynote speaker, *The Nature Conservancy Mid-Atlantic Region GIS Conference*, Conshohocken, PA, March 1998.

Invited speaker, *GIS for Brownfields Redevelopment*, Arizona Department of Environmental Quality, November 1996.

Organizer and speaker, *Brownfields and Beyond*, March 1996, New York City.

Invited speaker, *Statistics in Environmental Applications*, American Statistical Association conference held at the University of Delaware, April 1995.

Developer of the Government Institutes' two-day course on *Environmental Sampling*, Washington, D.C., October 1994, and Orlando, FL, January 1995.

Selected Publications

Guagliardo, Mark F., William A. Huber, Deborah M. Quint, and Stephen J. Teach, 2007. *Does Spatial Accessibility of Pharmacy Services Predict Compliance with Long Term Control Medications?* Journal of Asthma, 44:10, 881-883. doi: 10.1080/02770900701752680

Cox, LA and WA Huber, 2007. *Symmetry, Identifiability, and Prediction Uncertainties in Multistage Clonal Expansion (MSCE) Models of Carcinogenesis*. Risk Analysis 2007 Dec(6): 1441-53. doi: 10.1111/j.1539-6924.2007.00980.x

Sinton, Diana and William A. Huber, 2007. *Mapping Polka and Its Ethnic Heritage in the United States*. Journal of Geography **106** 41-47. doi: 10.1080/00221340701487913

Jamall, IS, T Lu, and WA Huber, 2005. *Distinguishing Between Multiple Chlorinated Solvent Plumes: A Comprehensive Approach*. The Annual International Conference on Soils, Sediments, and Water, Amherst, MA.

Cox, LA, D Babayev, and WA Huber, 2005. *Limitations of Qualitative Risk Assessment*. Risk Analysis **25** (3), 651-662. doi: 10.1111/j.1539-6924.2005.00615.x

Huber, William A., 2002. *GIS & Steganography—Part 3: Vector Steganography*. Published on the Web in Directions Magazine at http://www.directionsmag.com/article.php?article_id=195&trv=1 , April 18, 2002.

Huber, William A., 2001. *Estimating Markov Transitions*. Journal of Environmental Management, **v 61,** no. 4, pp 381-385. 10.1006/jema.2000.0412.

Huber, William A., 2000. *Variability and Uncertainty*. Chapter 12.2 of *The Standard Handbook of Environmental Science, Health, and Technology*, J. Lehr, Ed. McGraw-Hill.

Huber, William A. and W. A. S. Nijenuis, 2000. *Predictive Modeling of Ammonia Deposition from Large Numbers of Agricultural Sources* . 4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects and Research Needs. Banff, Alberta, Canada, September 3 - 8, 2000.

Huber, William A., 1999. *Convolution*. Published in three parts on the Web in Directions Magazine at http://www.directionsmag.com/features.asp, October 1999.

Harkness, Bracco, Franz, Tsentas, Becker, Huber, Orient, Rich, & Figura, 1998. *Natural Attenuation of Chlorinated Aliphatics at the Naval Air Engineering Station, Lakehurst,* NJ. In *Natural Attenuation/Chlorinated and Recalcitrant Compounds*, Wickramanayake & Hinchee, Eds.

Huber, William A, 1996. Discussion: *Detection of Low-level Environmental Pollutants.* Environmental and Ecological Statistics.

Huber, William A., and Douglas W. Watt, 1994. *Probabilistic Data Analysis and Soil Vacuum Extraction Used for Identifying the Location of DNAPLs.* Technical Papers of the Twelfth Annual Environmental Management and Technology Conference International, Philadelphia, PA, June 1994. Pages 492-513.

Huber, William A., 1993. Discussion: *Resampling from Stochastic Simulations for Assessing Uncertainty in Global Estimation*. Journal of Environmental Statistics, v. 1, no. 2.

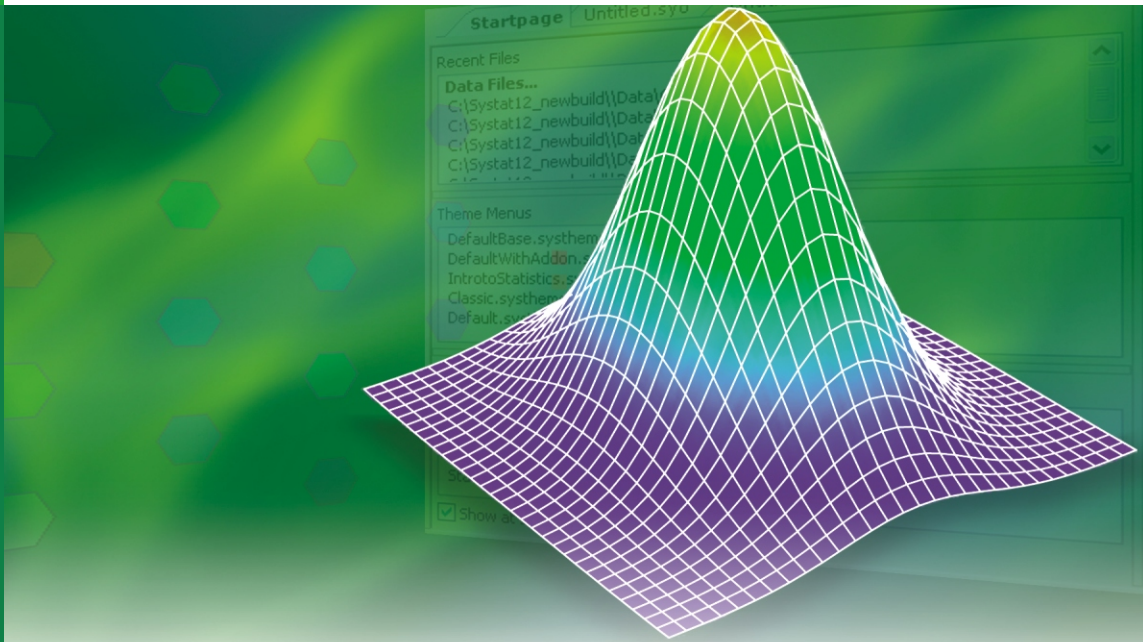
Huber, William A., 1993. *Graphical Techniques for Enhancing the Utility of Multivariate Environmental Statistics*. Multivariate Environmental Statistics, G.P. Patil et al., eds., North Holland/Elsevier, 1993. Pages 203-213.

Huber, William A., 1992. *Selecting a Statistical Methodology for RCRA Facilities*. Short Course, HMCRI Superfund '92, Washington, D.C.

Huber, William A., 1989. *Well Placement and Well Elimination*. NWWA conference on solving water problems with models, Indianapolis, Indiana. pages 187-207.

Huber, WA and C Bottcher, 1980. *Dielectronic Recombination in a Magnetic Field.* J. Phys. B: At. Mol. Phys. **13** L399-L404.

SYSTAT® 12



Statistics I II III IV

SYSTAT®

WWW.SYSTAT.COM

For more information about SYSTAT® software products, please visit our WWW site at *http://www.systat.com* or contact

Marketing Department
SYSTAT Software, Inc.
225 W. Washington Street, Ste. 425
Chicago, IL 60606
Phone: (877) 797-8280
Fax: (312) 220-0070
Email: info-usa@systat.com

1 2 3 4 5 6 7 8 9 0   05 04 03 02 01 00

Chapter
12

Factor Analysis

Herb Stenson and Leland Wilkinson

FACTOR provides principal components analysis and common factor analysis (maximum likelihood and iterated principal axis). SYSTAT has options to rotate, sort, plot, and save factor loadings. With the principal components method, you can also save the scores and coefficients. Orthogonal methods of rotation include varimax, equamax, quartimax, and orthomax. A direct oblimin method is also available for oblique rotation. Users can explore other rotations by interactively rotating a 3-D Quick Graph plot of the factor loadings. Various inferential statistics (for example, confidence intervals, standard errors, and chi-square tests) are provided, depending on the nature of the analysis that is run.

Resampling procedures are available in this feature.

Statistical Background

Principal components (PCA) and common factor (MLA for maximum likelihood and IPA for iterated principal axis) analyses are methods of decomposing a correlation or covariance matrix. Although principal components and common factor analyses are based on different mathematical models, they can be used on the same data and both usually produce similar results. Factor analysis is often used in exploratory data analysis to:

- Study the correlations of a large number of variables by grouping the variables in “factors” so that variables within each factor are more highly correlated with variables in that factor than with variables in other factors.
- Interpret each factor according to the meaning of the variables.

is to compute as many factors as there are eigenvalues greater than 1.0—so, in this run, you study results for two factors. After examining the output, you may want to specify a minimum eigenvalue or, very rarely, a lower limit.

Unrotated loadings (and orthogonally rotated loadings) are correlations of the variables with the principal components (factors). They are also the eigenvectors of the correlation matrix multiplied by the square roots of the corresponding eigenvalues. Usually these loadings are not useful for interpreting the factors. For some industrial applications, researchers prefer to examine the eigenvectors alone.

The *Variance explained* for each component is the eigenvalue for the factor. The first factor accounts for 58.4% of the variance; the second, 22.1%. The *Total Variance* is the sum of the diagonal elements of the correlation (or covariance) matrix. By summing the *Percent of Total Variance Explained* for the two factors ($58.411 + 22.137 = 80.548$), you can say that more than 80% of the variance of all eight variables is explained by the first two factors.

In the *Rotated Loading Matrix*, the rows of the display have been sorted, placing the loadings > 0.5 for factor 1 first, and so on. These are the coefficients of the factors after rotation, so notice that large values for the unrotated loadings are larger here and the small values are smaller. The sum of squares of these coefficients (for each factor or column) are printed below under the heading *Variance Explained by Rotated Components*. Together, the two rotated factors explain more than 80% of the variance. Factor analysis offers five types of rotation. Here, by default, the orthogonal varimax method is used.

To interpret each factor, look for variables with high loadings. The four variables that load highly on factor 1 can be said to measure "lankiness"; while the four that load highly on factor 2, "stockiness." Other data sets may include variables that do not load highly on any specific factor.

In the factor scree plot, the eigenvalues are plotted against their order (or associated component). Use this display to identify large values that separate well from smaller eigenvalues. This can help to identify a useful number of factors to retain. Scree is the rubble at the bottom of a cliff; the large retained roots are the cliff, and the deleted ones are the rubble.

The points in the factor loadings plot are variables, and the coordinates are the rotated loadings. Look for clusters of loadings at the extremes of the factors. The four variables at the right of the plot load highly on factor 1 and all reflect length. The variables at the top of the plot load highly on factor 2 and reflect width.